Satellite data efficient processing with dynamic block archive access

A. A. Proshin, A. M. Matveev, A. V. Kashnitskiy, M. A. Burtsev

Space Research Institute, Russian Academy of Sciences, Moscow, 117997, Russia

E-mail: andry@iki.rssi.ru

Abstract. The presently many process and phenomena research and monitoring tasks are solved with a broad range of various satellite data. One of the most actual and difficult problems is fusion processing of different satellite data types because of data heterogeneity. The paper describes the dynamic block access technology for various satellite data types for efficient processing, primarily from the computational resources and networking utilization perspective.

Accepted: 15.09.2020
DOI: 10.21046/2070-7401-2020-17-6-56-60

1. Introduction

Presently many process and phenomena research and monitoring tasks are solved with a broad range of various satellite data [1–3] having different spatial and temporal resolution, spectral bands etc. Generally, the storage scheme for any of satellite data types is optimized particularly for it in terms of both geographical projections for keeping the data and data tiling scheme. The data can be tiled in some regular way for a given projection or split into regular scenes, e. g. by division of a sensing orbit into equal sized overlapping fragments. Figure 1 shows contours of fragments used for keeping different types of data. TERRA/AQUA MODIS data, stored in sinusoidal projection regular granules, is shown in red. Data from the Russian KMSS instrument installed onboard the Meteor-M series satellites is shown in green. Data from Landsat-8 satellite is shown in orange. The depicted heterogeneity of satellite data really sophisticates the development of new data processing routines, and fusion data processing and analysis becomes even more sophisticated.

It is important to notice that many actual natural and anthropogenic factors monitoring require long term time series data compositing over large areas for various time spans, e. g. daily, weekly, monthly etc. Such kind of processing is usually consuming a lot of resources making parallel processing of fixed area of interest data fragments necessary. When all the source data is stored in the same regular tiled way as the aforementioned USGS MODIS data, parallel processing is quite simple. When you have to process the data without tiled storage scheme, like OLI/TIRS or KMSS, things get worse. When you have to fuse the data with different storage schemes, everything becomes very complicated.

Conventional data access schemes provide input files for processing in the same way as they are stored in the archives. Thus, processing for the given area of interest (AOI) requires all the data fragments intersecting it, even slightly. Usually such datasets are very excessive and result in additional network and disk overhead. Such excessiveness is shown on figure 1, where the required area of interest is marked with a square. Also, processing often requires only a subset of spectral bands stored in source data files. Even more, sometimes only downsampled data is required for processing. These factors can also contribute much to dataset excessiveness and overheads of the conventional access

scheme. Furthermore, processing of the archived files "as is" leads to necessity of preliminary area of interest data preparation stage during any processing routine. If fusion data processing is needed, this data preparation stage can also include reprojection of all the data to a chosen projection, resampling to a chosen resolution, spectral bands selection etc., complicating the processing more and more.

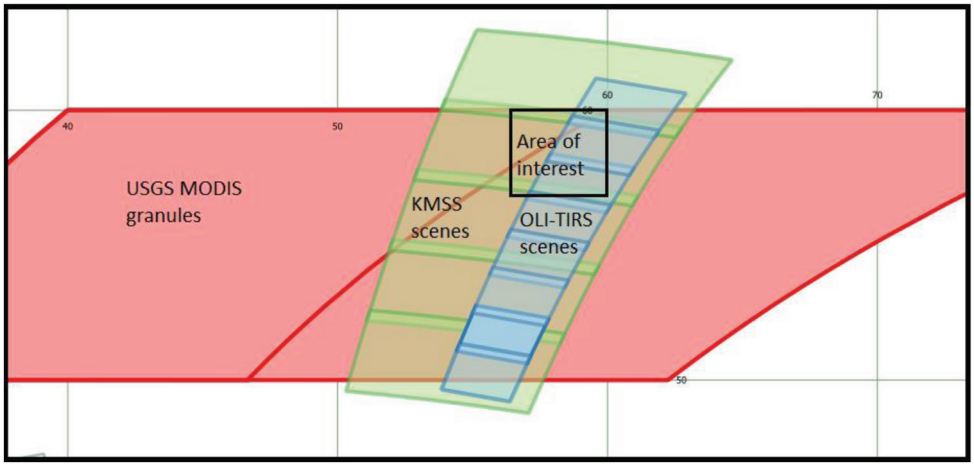

Figure 1. Contours of different satellite data storage schemes.

Obviously, these overheads can be avoided by providing only the required data fragments to processing routines. It can be achieved with regular data tiling storage and access scheme, addressed further as "block access", significantly simplifying the processing routines development and implementation.

One of potential ways to implement the proposed scheme is restructuring of archives to store various satellite data uniformly projected and tiled. It may seem the most obvious and efficient way, but actually it has a lot of drawbacks. First, any reprojection usually reduces the measurements precision. Second, there is no tiling scheme equally efficient for different types of satellite data and processing tasks and keeping the same data in several tiling schemes simultaneously results in great storage overheads. Thus, the most practical choice is dynamic data preparation for required tiling scheme, i. e. dynamic block access.

Processing of large satellite data volumes requires high computational, storage and networking performance making the overhead minimization for dynamic block access the most important. One of the most efficient ways of data preparation for tiling is direct access to archived files, i. e. the option of reading only the requested part of the file, both for area of interest and spectral bands, and then send it over the network. Of course, this option must be supported by the file representation format. The most common file format with georeference support is GeoTIFF, also supporting an internal tile structure. Another important overhead optimization factor is the optimal tile size selection defining file read and network transfer speeds. The paper describes the dynamic block access technology for satellite data efficient processing developed in IKI. Further, the internal storage tile will be referred to as tile, logical tile for processing will be referred to as block, a number of blocks for processing will be referred to as blockset.

2. Technology description

The described block access scheme is inseparable from distributed parallel data processing. In IKI framework [4, 5] processing is managed by the dedicated dispatch server composing datasets for processing on remote nodes. The processing nodes have a universal software set installed for a wide range of different processing types. Each dataset for processing is composed of a set of instructions for processing and a set of data files locations. Any free processing node receives a next processing task in accordance with centrally-controlled processing priorities, performs it and sends the results to the archive location, also defined in the processing dataset.

When processing nodes have direct access to files in the storage, i. e. over NFS protocol, the data preparation routine for selected tiling scheme can be ran on them. More often the archives are

geographically distributed and direct access to their contents is impossible. In this case dedicated data tiling services balanced on many hardware instances in each data storage center are necessary.

The dynamic block access problem becomes even more complicated when the processing algorithm requires data both for the area of interest and the AOI border areas (extended AOI). If we have to process a large area split into several adjacent blocksets, same data will be processed several times due to extended blocksets intersection. Network and service performance optimization in this case is achieved by implementing a special data preparation stage for caching all the blocks in the dedicated, high-performance disk buffer. During the dataset preparation in this case the AOI is split into overlapping rectangular blocksets including blocks both for required area and for border ones. E.g., if the processing area is split into 10 by 10 blocksets, the 12 by 12 blocksets will be provided from fast cache. The preliminary data preparation stage can also minimize the data download time for processing nodes, thus maximizing their computing time. Preliminary preparation with caching is extremely efficient for numerous processing of the same dataset for different derived products. After all the processing is done, the cache buffer is flushed.

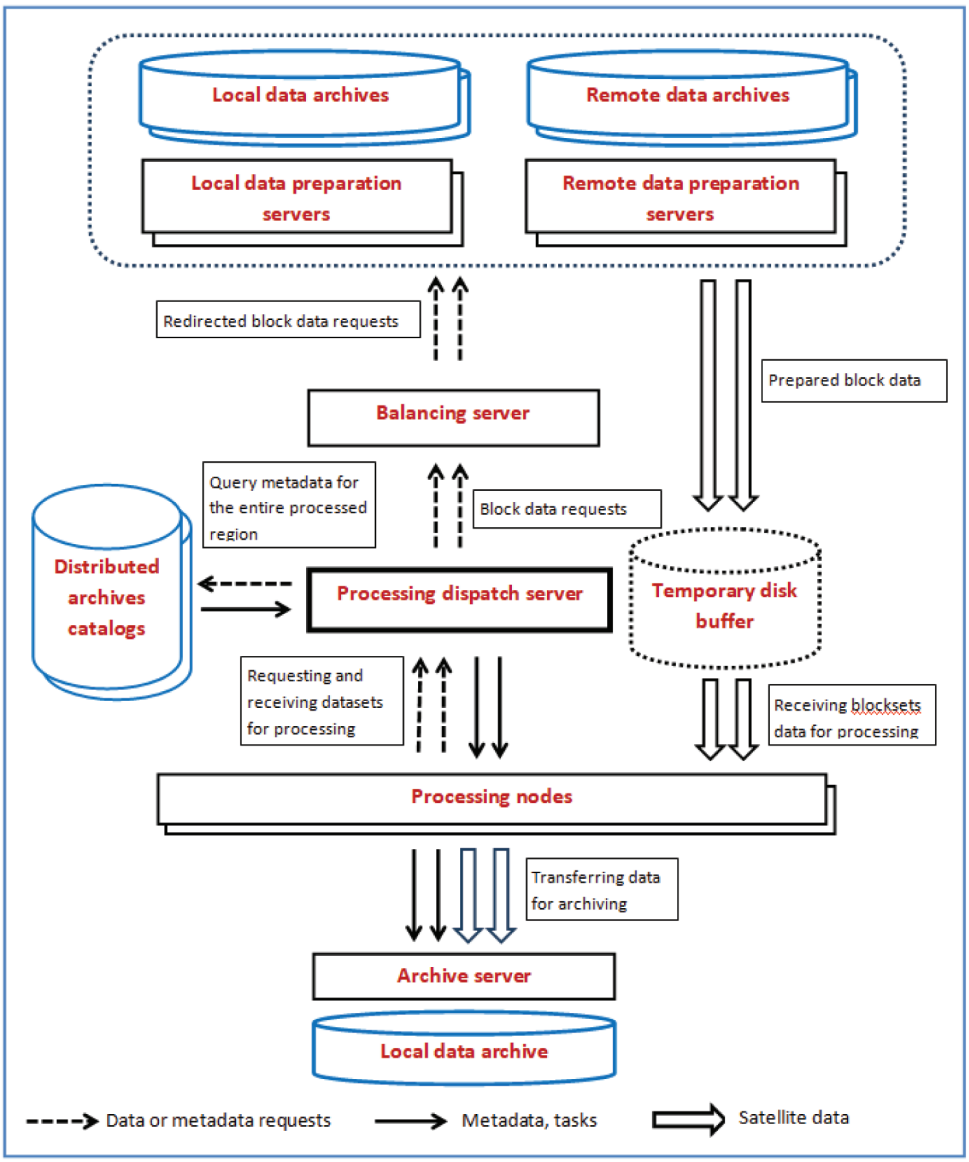

Figure 2. Functional diagram of satellite data processing with dynamic data block access.

Figure 2 depicts the functional diagram of satellite data processing with dynamic data block access. The dispatch server, placed in the diagram center, manages the processing dataset composition. This procedure includes:

- Reading the processing configuration.
- Querying the processed data general information.
- Preparing queries to the data preparation services for required blocks and data locations.
- Sending the prepared queries via the Balancing server.
- Accumulating the selected blocks in the cache buffer.
- Composing the dataset.

Balancing server hosts a dedicated load balancing daemon redirecting queries to data preparation services. These services can be hosted both on data storage servers and dedicated nodes with high-speed access to archives. The load is flexibly balanced in accordance to data locations, each service instance load limits etc.

The composed datasets are received by processing nodes, the required blocks are downloaded there and processed. The processing results and corresponding metadata are sent to local archive.

An overhead analysis was performed for all stages of processing from preliminary data preparation to results archiving for efficiency optimization. It showed that two main contributors for overheads are AOI block tiling optimality (i. e. block size) and blockset size. Reducing the block size significantly increases the preparation and networking overheads along with reducing the file reading speed. On the other hand, enlarging the block size significantly increases the amount of border area for a single blockset. Increasing the blockset size can minimize the border area overheads, but it leads to processing time growth and paralleling capacity decrease. Generally, the optimal block and blockset sizes depend on a great variety of factors including data type, processing routine specifics, available hardware, acceptable processing time etc. At this moment the authors are developing a method for optimal block and blockset parameter selection for processing routines development and implementation.

3. Implementation summary

Implementation of the presented dynamic block access scheme is based on the IKI-Monitoring shared use center (SUC) [6, 7] software and hardware infrastructure and technologies developed by IKI. SUC archives are built with the UNISAT technology [8] for very large distributed satellite data archives. Satellite data files are stored in the GeoTIFF file format with internal tiling support. Data processing system is built with the distributed parallel satellite data processing technology [5] for efficient management of multiple processing nodes.

The following software elements were developed specifically for the scheme implementation:

- Data block preparation GCI-utility, based on GDAL software.
- Redirection and balancing daemon for data preparation services.
- Blocks query compositing for various satellite data types software.
- Multithreaded data download software based on aria2c open source downloader.
- Processing dataset compositing routine, adapted for block access.

All the software is written in Perl and Python languages and operates under the UNIX-based OS-es.

4. Conclusions

The described technology was implemented for cloud-free compositing of the Russian Meteor-M satellite KMSS instrument data. This compositing routine uses OLI/TIRS data for georeference adjustment and atmospheric corrected MODIS data for calibration enhancement. Dynamic block access have significantly simplified the data fusion implementation and increased the processing speed. Migration of other processing routines to this scheme is on the way along with the method of optimal block and blockset size selection development.

Acknowledgements

The work is performed in the frame of "Big data in space research: astrophysics, Solar system, geosphere" Research Theme (state reg. No. 0024-2019-0014).

5. References

[1] Loupian E. A., Bourtsev M. A., Proshin A. A., Kobets D. A., 2018 Evolution of remote monitoring information systems development concepts, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, Vol. 15(3), pp. 53–66, DOI: 10.21046/2070-7401-2018-15-3-53-66.

[2] *Satellites to be built and launched by 2026*, 2017, p. 7, available at: URL: http://www.euroconsult-ec.com/research/satellites-built-launched-by-2026-brochure.pdf.

[3] Zhu L., Suomalainen J., Liu J., Hyyppä J., Kaartinen H., Haggren H., *A Review: Remote Sensing Sensors*, IntecOpen, 2018, DOI: 10.5772/intechopen.71049.

[4] Kobets D. A., Matveev A. M., Proshin A. A., Mazurov A. A., Operation control and management of distributed complexes of automatic streaming processing of satellite data, *15th Int. Scientific and Technical Conf. "Actual problems of creation of space remote sensing systems of the Earth"*, *Electromechanical matters*, *Proc.*, 2018, VNIIEM studies, 2018, pp. 225–234.

[5] Kobets D. A., Matveev A. M., Mazurov A. A., Proshin A. A., Organization of automated multithreaded processing of satellite information in remote monitoring systems, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2015, Vol. 12(1), pp. 145–155.

[6] Loupian E. A., Proshin A. A., Bourtsev M. A., Kashnitskii A. V., Balashov I. V., Bartalev S. A., Konstantinova A. M., Kobets D. A., Mazurov A. A., Marchenkov V. V., Matveev A. M., Radchenko M. V., Sychugov I. G., Tolpin V. A., Uvarov I. A., Experience of development and operation of the IKI-Monitoring center for collective use of systems for archiving, processing and analyzing satellite data, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2019, 2019, Vol. 16(3), pp. 151–170, DOI: 10.21046/2070-7401-2019-16-3-151-170.

[7] Loupian E. A., Proshin A. A., Bourtsev M. A., Balashov I. V., Bartalev S. A., Efremov V. Yu., Kashnitskiy A. V., Mazurov A. A., Matveev A. M., Sudneva O. A., Sychugov I. G., Tolpin V. A., Uvarov I. A., IKI center for collective use of satellite data archiving, processing and analysis systems aimed at solving the problems of environmental study and monitoring, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2015, Vol. 12(5), pp. 263–284.

[8] Proshin A. A., Loupian E. A., Balashov I. V., Kashnitskiy A. V., Bourtsev M. A., Unified satellite data archive management platform for remote monitoring systems development, *Sovremennye problemy distantsionnogo zondirovaniya Zemli iz kosmosa*, 2016, Vol. 13(3), pp. 9–27, DOI: 10.21046/2070-7401-2016-13-3-9-27.